# 11

# Gaining Credibility Through Root-Cause Analysis and Exception Handling

*Plurality which is not reduced to unity is confusion; unity which does not depend on plurality is tyranny*.

BLAISE PASCAL (1623–1662)

---

This chapter introduces several analytical tools that are useful as regression model diagnostics and in situations where normality of the error distribution cannot be assumed. This chapter describes

- what role correlations and residuals play in validating modeling assumptions
- how elasticities, both own-price and cross-elasticity, are essential for understanding business growth as well as revenue-quantity relationships
- how a resistant measure of correlation can safeguard making unwarranted statements about causation
- why nonconventional methods are so important in improving the robustness of forecasting models in the face of uncertainty
- why forecast error patterns can give valuable insights into improving forecasting performance

## *The Diagnostic Checking Process in Demand Forecasting*

Many demand planners and forecasters proceed by seeing how alike things are. Others proceed by trying to understand why things are different. The root-cause analysis of residuals and forecast errors is consistent with the latter approach. The diagnostic checking process is designed to reveal departures from assumptions about the underlying distribution of random errors and model formulation. It is an important phase for demand forecasters to learn about, as it can be a powerful visual tool for assessing the potential usefulness of a forecasting model, isolating and correcting unusual values, identifying hidden patterns, improving forecast accuracy, and understanding the nature or randomness in time series data.

A residual analysis may suggest nonlinear relationships, the need for transformations, or a better understanding of patterns and events that may not be transparent in the bulk of the unexplored data. Forecast error patterns can give valuable insights into improving forecast accuracy. And as we shall see, the unplanned findings of correlation analysis in regression modeling will often yield the most interesting and important results in a demand forecasting application.

## *The Role of Correlation Analysis in Regression Modeling*

When demand forecasters have reason to believe that more than one factor (driver of demand) is required to solve a demand forecasting or econometric analysis problem, they turn to multiple linear regression models. To make forecasts with a multiple linear regression model, one needs to provide forecasts of more than one independent (explanatory) variable. When multiple variables are involved, the modeling effort can easily become more involved, so it is necessary to start with exploratory data analysis tools to get proper insight into the forecast-generating process.

By focusing only on the fit of the equation, you may discover a useful description of a forecast profile (*change*) but at the same time misspecify the uncertainty (*chance*), because error assumptions are needed in using the equation for forecasting. Also, accurate forecasts of the drivers of demand are needed, and if the underlying model errors are not normally distributed, as is usually assumed, there is a need for identifying root causes and exceptions. So there is always a trade-off between the goodness of fit and the quality of fit in the use of regression models for forecasting applications.

Recall, In a **multiple linear regression** (MLR) analysis, the deterministic component (*change*) takes the form of an equation,

$$\mu_{Y(X)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where $X_1, \ldots, X_k$ are $k$ independent variables (or regressors), and $\beta_0, \beta_1, \ldots, \beta_k$ are called **regression parameters**. This regression equation arises when the variation in the dependent variable $Y$ is assumed to be affected by changes in more than one independent variable. Thus, the *expected (or* average*) value* of $Y$ is said to depend on $X_1, X_2, \ldots, X_k$. The dependence on $X$ is henceforth suppressed in the notation; Let $\mu_{Y(X)} = \mu_Y$. In this case, one speaks of a multiple linear regression of $Y$ on $X_1 \ldots X_k$.

The regression equation is called a **model** when a random error (*chance*) term is added to the equation, which then becomes $Y = \mu_{Y(X)} + \varepsilon$, where $\varepsilon$ commonly has an assumed normal error distribution.

While the formal theory of normal multiple linear regression analysis is extensive and is dealt with in many business statistics textbooks, of interest here are its application and interpretation in demand forecasting problems and those derivations and algorithms that are directly applicable to the interpretation of the analysis.

Adequacy of the model assumptions can be examined through a variety of methods, frequently graphical and mostly involving residuals (Actual minus Fit). One must be aware of a range of regression pitfalls to be avoided. These include trend collinearity, overfitting, extrapolation, outliers, nonnormal distributions, multicollinearity, and invalid assumptions regarding the model errors (e.g., independence, constant variance, and, usually, normality). Such inadequate assumptions often point to the root causes of not-so-credible forecasts.

But how does one put a model together? The first step in beginning a regression analysis for demand forecasting is to identify the drivers of demand—called factors—independent or explanatory variables that are believed to have influenced and expected to continue to influence the (dependent) variable to be forecast. The scatter diagram is a useful graphical tool for exploring the relationships among such factors.

The second step is to create a regression model by estimating the coefficients in the model by the method of least squares. This will give us a fitted equation from which we can determine the forecast profile.

## Linear Association and Correlation

When the values of one time series (or variable) are paired with corresponding values of a related time series (or variable), a relationship between the variables can be depicted in a scatter diagram with one variable is plotted on the horizontal scale and the other is plotted on the vertical scale. Such a plot is a valuable tool for studying the relationship between a pair of variables.

When two series have a strong positive association, the scatter diagram reveals a pattern of points along a line of positive slope. A negative association shows up as a scatter pattern along a line with negative slope. A conventional measure of such linear association between a pair of variables $Y$ and $X$ is given by the *Pearson* **product moment correlation coefficient** $r$, where $r$ is an averaging formula using the sample mean and sample standard deviation of the two variables, respectively.

***The product moment correlation coefficient is a conventional measure of linear association between two variables.***

Although forecasting and statistical software programs routinely calculate $r$, it is useful to view it as the result of an averaging process; namely, of the average of a product of standardized variables: Average {(Standardized $Y_t$)*(Standardized $X_t$)}, with a divisor of ($n$–1) instead of $n$, where $n$ is the common number of $X$, $Y$ pairs. When a variable is standardized, it has a zero mean and a unit standard deviation, which is useful for making comparisons and correlations between variables that have very different sizes or scales of measurement.

***A standardized value is obtained by subtracting the sample mean from the data and dividing by the sample standard deviation.***

Figure 11.1 shows a spreadsheet calculation for obtaining the product moment correlation between annual housing starts and mortgage rates. The coefficient can vary between +1 and –1, so that $r$ = –0.20 suggests a weak negative association between housing starts and mortgage rates.

Although the product moment correlation coefficient for housing starts versus mortgage rates data shown in the figure is only about –0.20, the correlation coefficient for the respective annual change in these variables turns out to be –0.57. Both are negative, as expected, but the latter reflects a much stronger linear association. This suggests that the strength of the relationship between housing starts and mortgage rates is reflected in their respective growth rates, not so much the actual levels.

## The Scatter Plot Matrix

Creating scatter diagrams to validate a **linear association** between the dependent variable and each of the independent variables, as well as between pairs of independent variables, is an essential step in exploratory data analysis for larger datasets. Doing so can save you time when questions arise about root causes and exceptions with flawed models. At the same time, the diagrams can provide a better understanding of the data-generating process in the underlying relationships

| Housing Starts | Mortgage Rates | Col 1 × Col 2 |
|---|---|---|
| 0.168898 | −1.384396 | −0.233822 |
| −0.051812 | −1.405769 | 0.072836 |
| −0.217938 | −1.410043 | 0.307302 |
| −1.131335 | −1.239058 | 1.401789 |
| −0.755474 | −1.157839 | 0.874718 |
| −0.114703 | −0.944107 | 0.108292 |
| −0.235737 | −0.589311 | 0.138923 |
| −0.334226 | −0.328558 | 0.109813 |
| 1.500873 | −0.619234 | −0.929391 |
| 2.403886 | −0.679079 | −1.632429 |
| 1.480404 | −0.538016 | −0.796480 |
| −0.618717 | −0.140474 | 0.086913 |
| −1.144684 | −0.123375 | 0.141226 |
| −0.026003 | −0.114826 | 0.002986 |
| 1.307751 | −0.102002 | −0.133393 |
| 1.406240 | 0.124554 | 0.175153 |
| 0.589850 | 0.620413 | 0.365951 |
| −0.753694 | 1.377025 | −1.037857 |
| −1.370734 | 2.176384 | −2.983244 |
| −1.435998 | 2.330271 | −3.346265 |
| 0.464958 | 1.312906 | 0.610447 |
| 0.602902 | 1.214589 | 0.732279 |
| 0.580060 | 0.876892 | 0.508650 |
| 0.768732 | 0.299815 | 0.230477 |
| 0.220219 | −0.042157 | −0.009284 |
| −0.172550 | −0.089178 | 0.015388 |
| −0.504802 | 0.312639 | −0.157821 |
| −1.048865 | 0.274167 | −0.287564 |
| −1.577502 | −0.012234 | 0.019300 |
| 0.000000 | 0.000000 | $r = -0.201611$ |
| 1.000000 | 1.000000 | |

Figure 11.1 Calculation of the product moment correlation coefficient between annual housing starts and mortgage rates. (*Source*: Figure 7.12 Housing starts data; Figure 7.8 Mortgage rates data**)**

An arrangement of scatter diagrams between multiple pairs of variables is called the **scatter plot matrix**. We can summarize this by creating a correlation matrix of **product moment correlations** between pairs of variables.

The diagonal of a correlation matrix consists of 1s because each variable is perfectly correlated with itself. At the intersection of each row and column is the correlation coefficient relating the row variable to the column variable.

Because the matrix is symmetrical, it is useful to display the product moment correlation coefficient r on one side of the diagonal and an outlier-resistant version of correlation called r* (SSD)—defined in the next section—on the other. To distinguish it from other types of correlation measures, the term ordinary correlation coefficient is here used interchangeably with the term Pearson product moment correlation coefficient.

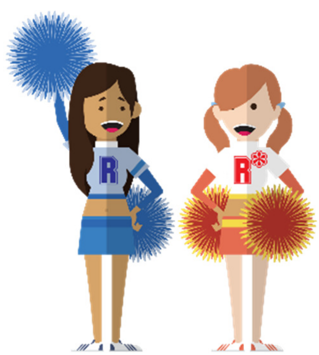
In this way, the demand forecaster gets the necessary insight from the augmented correlation matrix for constructing regression relationships. Contrasting r with r* (SSD) may indicate departures from linearity due to outlying or non-typical data, so the demand forecaster then needs to review the underlying data for nonlinearity in the patterns. An outlier may not necessarily appear visually extreme from the bulk of the data in these situations.

***A scatter plot matrix is an array of associations between pairs of variables.***

## The Need for Outlier Resistance in Correlation Analysis

The robust estimator of correlation, known as $r^*$(SSD), is less sensitive to outliers than the ordinary correlation coefficient *r.* It is derived from the standardized sums and differences of two variables, say *Y* and *X*, as introduced in a 1975 *Biometrika* paper entitled "*Robust Estimation and Outlier Detection with Correlation Coefficients*," by Susan J. Devlin, Ram Gnanadesikan, and Jon R. Kettenring.

The first step in obtaining $r^*$(SSD) is to standardize both *Y* and *X* robustly by constructing two new variables *Y* and *X*:




$$\bar{Y} = (Y - Y^*)/S_Y^* \text{ and } \ddot{X} = (X - X^*)/S_X^*$$

where $Y^*$ and $X^*$ are robust/resistant estimates of location and $S_Y^*$ and $S_X^*$ are robust/resistant estimates of scale.

Now, let $Z_1 = \bar{Y} + \ddot{X}$ and $Z_2 = \bar{Y} - \ddot{X}$, the sum and differences vectors, respectively. Then the robust variance of the sum vector $Z_1$ and difference vector $Z_2$ are calculated; they are denoted by $V_+^*$ and $V_-^*$, respectively. These variances are used in the calculation of the robust correlation estimate $r^*$(SSD) given by

$$r^*(\text{SSD}) = (V_+^* - V_-^*) / (V_+^* + V_-^*).$$

The justification for this formula can be seen by inspecting the formula for the variance of the sum of two variables:

$$\text{Var}(Z_1) = \text{Var}(\bar{Y}) + \text{Var}(\ddot{X}) + 2 \text{ Cov}(\bar{Y}, \ddot{X})$$

where Cov denotes the covariance between $\bar{Y}$ and $\ddot{X}$.

Because *Y* and *X* are standardized, centered about zero, with unit scale, the expected variance of $Z_1$ is approximately

$$\text{Var}(Z_1) \approx 1 + 1 + 2\,\rho\,(\bar{Y}, \ddot{X}) = 2\,(1 + \rho)$$

where ρ is the theoretical correlation between $\hat{Y}$ and *X*.

Similarly, for $Z_2$,

$$\mathrm{Var}\ (Z_2) \approx \mathrm{Var}\ (\bar{Y}) + \mathrm{Var}\ (\ddot{X}) - 2\ \mathrm{Cov}\ (\bar{Y}, \ddot{X})$$

$$= 1 + 1 - 2\ \rho\ (\bar{Y}, \ddot{X}) = 2\ (1 - \rho).$$

Notice that ρ is given by the expression

$$[\mathrm{Var}\ (Z_1) - \mathrm{Var}\ (Z_2)]\ /\ [\mathrm{Var}\ (Z_1) + \mathrm{Var}\ (Z_2)] \approx [2(1+ \rho) - 2(1- \rho)]\ /\ [2(1+ \rho) + 2(1 - \rho)] = \rho.$$

Some robust estimates of the (square root of the) variance, required in the formula for $r^*$ were discussed in Chapter 2; these include the unbiased median absolute deviation from the median (UMdAD =MdAD/0.6745) and the unbiased interquartile difference (UIQD = IQD/1.349).

## *Using Elasticities*

In practice, it turns out that looking at correlations and linear relationships among transformed values of variables in a revenue demand model can also be very useful. For example, when using elasticities, the period-to-period changes (month-to-month, year-over-year, etc.) and percentage changes are often applied to time series modeling for a root-cause analysis.

Two important determinants of a firm's profitability—indeed its survival—are cost and the demand for its products or services. Demand must exist or be created if the business is to survive. It must also be high enough at least to cover fixed costs. Because of its key role, all business-planning activities require a careful analysis of demand over time. Demand forecasters also need to be aware of the relationship between quantity demanded and price.

***Demand forecasters play an important role in helping to make pricing decisions by estimating price elasticities for products and services with regression models.***

**The complete chapter can be found in**

# Change & Chance Embraced

## ACHIEVING AGILITY WITH DEMAND FORECASTING IN THE SUPPLY CHAIN

HANS LEVENBACH, PhD



ISBN-10:0692945989
ISBN-13:978-069294588

Cover art: Rose Tannenbaum, Berkshire TypeGraphic, www.tgo.com

Cartoon art: Don Engel, Isographic LLC

Copy editor: David Coen, Freelance copyeditor

Printed by Createspace, www.createspace.com



Professional development: Certified Professional Demand Forecaster (CPDF®), www.cpdftraining.org

Contents

: